

# Living on Fumes: what “small data” interviews can tell us about Location-Based Services and Big Data’s increasing role in what we know and where we are.

Jim Thatcher\*

\* Clark University, Department of Geography

**Abstract.** The past year has seen a rise in the profile of ‘Big Data’ in both the public and private sectors. This paper examines the generation and use of spatial Big Data through Location-Based Services and Social Networks. Drawing on interviews of LBS/LBSN designers and developers, the paper introduces the concept of Data Fumes – where one application leverages the data produced by another. The reliance upon Data Fumes is shown to have profound effects on what can be known and done with spatial Big Data. Finally, the paper suggests spatial Big Data should be examined with an eye to the earlier debate of “systems vs. science” within the GIS community.

**Keywords:** Big Data, LBS, LBSN

## 1. Introduction

In late January of 2012, Daniel Rasmus, a writer for *Fast Company*, predicted 2012 to be the “year of Big Data” (Rasmus 2012). From marketing (Baker 2012) to health care (Cerrato 2012) to the national science funding agencies (NSF Press Release 2012), Big Data and its related models and analytical approaches have risen drastically in prominence. This brief paper examines the generation of spatial Big Data through the use of Location-Based Services (LBS) and Social Networks (LBSN). It proceeds in three parts: First, drawing from ethnographic interviews of mobile application designers and developers the term “Data Fumes” is introduced and defined as the predominant orientation of private companies creating LBS/LBSNs. Second, two effects this orientation has upon the data produced are demonstrated. On the one hand, both what information Data Fumes contain and who has access to them is controlled by an extremely small set of designers working for private corporations whose decisions promulgate through the

mobile application ecosystem. On the other, the “location” produced in spatial Big Data sets is separated from physical location and commoditized. The data generated through LBS and LBSN use is driven by a motive for profit both from the end-user and corporation, this motivation shapes the very definition and understanding of “location.” Finally, the paper concludes by drawing parallels between concerns in spatial Big Data research and the earlier “Systems versus Science” debates in the GIS community. Current research into LBS, LBSNs and Big Data more broadly will be well served by paying heed to these earlier discussions and research.<sup>1</sup>

## 2. Spatial Big Data and Data Fumes

Before turning to the concept of Data Fumes, it’s necessary to briefly define how Big Data is understood with respect to spatial information in this paper. From a technical perspective, data has always been big (Farmer and Pozdnoukhov 2012). As such “Big Data” presents an ever-shifting target that has less to do with size and more with the ability to rapidly combine, aggregate, and analyze diverse and disparate sets of information. Following Jacobs (2009, 39), the “pathologies of big data are primarily those of analysis.” Beyond the technical, the rapid growth of Big Data approaches and the acceptance of the Big Data movement in both private firms and academic research belies a socio-technical phenomenon that accepts certain epistemological and cultural views of scientific praxis and the nature of reality. In this sense, Big Data represents the latest iteration of the desire to find efficiency and meaning in quantitative analysis. Big Data requires a belief that life can be captured and modeled by data or even fully transformed into it (boyd and Crawford 2012; Berry 2011). With reality accurately captured in ever larger data-sets, scientific praxis transforms into a new “fourth paradigm” of manipulation and exploration (Hey et al. 2009). Large numbers of potential correlations are equally considered, in contrast to traditional methods driven from a theoretically based hypothesis and a small number of testable variables (Batty 2012). Science in the Big Data era is an abductive process where “hypotheses are developed to account for observed data” (Farmer and Pozdnoukhov 2012, 5).

The “Big Data perspective” for spatial information involves “using large-scale mobile data as input to characterize and understand real-life phenom-

---

<sup>1</sup> A significantly expanded version of this paper addressing academic study of spatial Big Data is presently under review at the *International Journal of Communication*.

ena” (Laurila et al. 2012, 1). At present, this mobile data comes from two predominant sources – *Twitter* and *Foursquare* – with the data captured and stored from mobile users of these applications is seen as containing “rich information” (Long et al. 2012). The Big Data viewpoint posits that check-ins and tweets reveal meaningful information that can be used by researchers to study society and by companies to increase profits. These two distinct goals are both found within the data already generated by end-users of certain applications, something referred to here as “Data Fumes.”

Since 2009, more than \$115 million has been invested in location based start-up companies, making them a major factor in the generation, manipulation and access of spatial information data sets (Wilson 2012). Many of these start ups have focused around what can be called the “check-in.” For the purposes of this paper, a “check-in” is defined as an action the end-user takes which broadcasts their supposed location at a given time and place. It is a purposeful, end-user initiated action which creates this data-point: “Users *check-in* at *venues* where they are present, effectively reporting their location” via the application to those who have access to the information – a group which may include friends, the makers of the application, other corporations, and researchers (Carbunar and Portharaju 2012). Check-ins may be distinguished from other more ubiquitous types of spatial tracking in that they are discrete events: a user checks-in at a movie theater and later checks-in at a restaurant, but the period between check-ins is not captured. Although other services exist, the dominant systems for checking-in at the moment are *Foursquare* and *Facebook places*. Although different sources cite different numbers, since 2009 it is estimated that *Foursquare* has seen over two and a half billion check-ins (Benner and Robles 2012). Meanwhile, *Facebook*, with over two-hundred million active users, has had upwards of two billion actions tagged with locations in April of 2012 alone (Long et al. 2012).

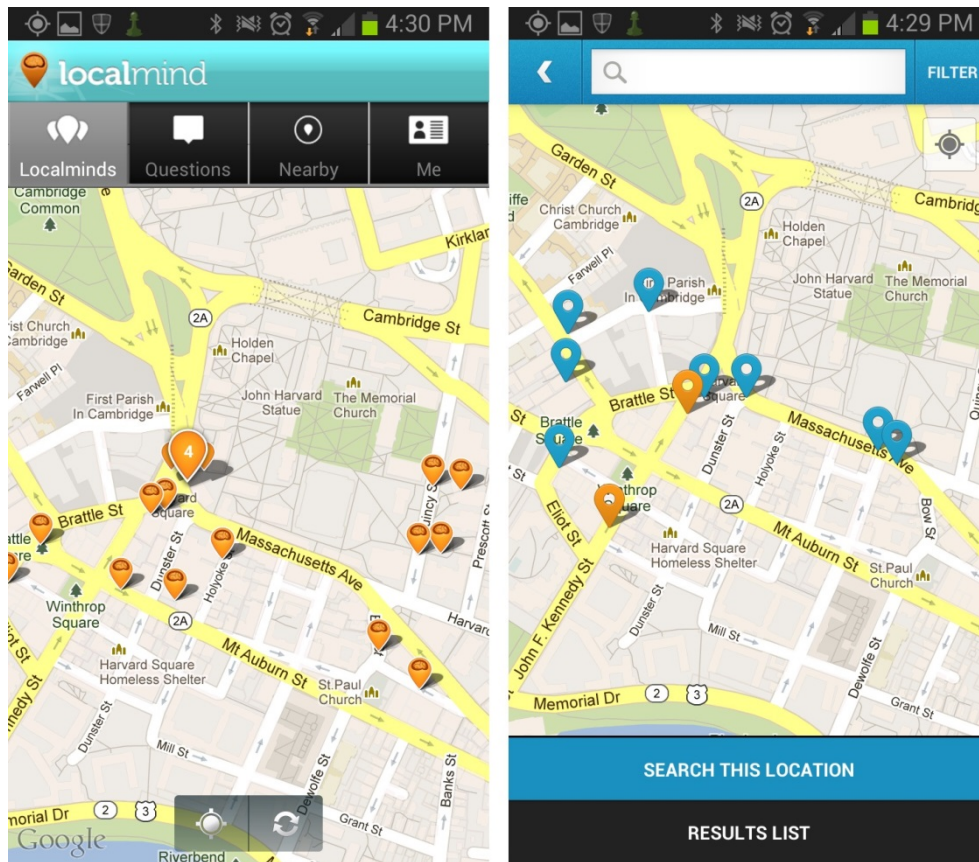
Data fumes are attempts to “add value” to the check-in, to make it more meaningful and to profit from the provision of this meaning. The term itself comes from an interview with the CEO of a popular mobile spatial start-up:

*John (names have been obfuscated):* In the end, [*his applications*] all tie to location data, they’re all sitting on top of other people’s location data and behaviors that we already have when we’re going out, like checking-in...

*John:* Say *Foursquare* didn’t exist, I’d have to convince everyone in the world to check-in and share their location. The fact that *Foursquare* exists and aims to get everyone in the world to check-in ... it means they can concentrate on that, and I can use all the data they’re already collecting, the

fumes, the exhaust of their data to do really interesting things.

Data fumes refer directly to the data and information that is generated through actions the end-user already takes. An application which “live[s] on data fumes,” as another interviewee, *Paul*, stated, is one which seeks to access and manipulate this already existing data in such a way as to “add value” to the end-user’s experience. As *John* put it, to maximize the “cost versus reward ratio.” In so doing, this prefigures a relationship with existing spatial data and representations. Of the mobile spatial applications represented in the interview set, only one was not building their product off of an existing platform. *John*’s most recent application, for example, uses *Foursquare* check-in data displayed on a map created by another third-party vendor. In total, every application relied on mapping information provided by either *Google* or *OpenStreetMap*. See figure 1 for an example of *Local Mind*’s use of *Foursquare*’s check-in locations with both applications using *Google* as their base layer on *Android* platform phones. Every interviewee had an acceptance that the data being generated, the millions of discrete check-ins, represented meaningful information on which to act. The existing behavior offered a rich source on which to “add value,” to capitalize on the Big Data sets already being generated. The next section of this article presents the two entwined problems with a reliance upon Data Fumes.



**Figure 1.** *Local Mind* on the left and *FourSquare* on the right, both use Google's basemap.

### 3. Access and Control

Data Fumes engage Big Data datasets with an inscription into and belief in the meaning of the data “given off” by actions end-users are already taken. This data is accepted as useful, meaningful, and ‘Big.’ These beliefs produce two entwined problems with the very nature of mobile Big Data: First, while the data sets may be big in size, their format, composition, and access is regulated by a very small number of individuals. Using information generated by other applications forces a reliance upon the means through which that data is made available. What can be known and what can be done with mobile spatial Big Data is delimited by the decisions of a very small set of programmers and businesses, their choices promulgating through the mo-

bile ecosystem. Second, as these decisions are made by private corporations engaged in profit seeking behavior, “location” itself has become a commodity able to be bid for, bought, and sold, but with little relation to actual physical location. This section demonstrates the entwined nature of these problems, tracing the limits placed on Big Data creation and access back to a small number of for-profit corporations and their developers.

Much of the allure of Data Fumes is that they build from data that already exists, that is already ‘given off.’ It lets start-ups focus on “add[ing] value based on work you’re already doing” (*John*). However, both the form of the information accessed and the ability to access it are held entirely outside the control of the vast majority of individuals. Despite the *size* of the data set, its *quality* and *characteristics* are entirely outside the control of the mobile start-up. When a start-up builds their application using *Foursquare* check-in data, they are accepting that their users will access information through the venue-based dataset that structures the *Foursquare* check-in.

For private companies, the lack of control and guaranteed access to information presents a set of concerns. For example, in May of 2012, *Foursquare* announced a change in their public API in order to prevent applications from accessing the information of users checked into venues other than that of the end-user (Thompson 2012). While this change was meant to prevent the creation of applications that could be used for stalking other users, like *Girls Around Me* (Brownlee 2012), the change in access to information affected a host of unrelated applications. One interviewee had to abandon their project and begin a new one as there was no way to continue their current application. *Assisted Serendipity*, an application previously praised by the *Foursquare* CEO, was left unable to function and ceased development (Thompson 2012). Private companies run clear risks of shifting access to the information they require when relying upon the data given off by interests other than their own. *Foursquare* changed their API policy due to bad publicity, *Twitter* drastically restricted free access to their data when they found a market willing to pay for it; in each case, the decision to change was made by a single corporation, but the repercussions were felt along the entire data chain. What can be known, what can be done, and what can be shown are inherently controlled by those who control access to the data.

In addition to basic concerns over access, reliance upon data given off by other corporations accepts an underlying view of individual and society shaped through the incentives that drive data creation and the imperatives of the corporations who drive it. The use of *Twitter* and *Foursquare* is built around a for-profit model in which end-users are incentivized to participate and the information they provide is commoditized and sold to advertisers.

This has a double effect on the commoditization of data produced. On one end, end-users may manipulate the system providing false information in order to receive rewards. On the other, the organization and form of the data produced is structured in such a way as to standardize and quantify “location” in order to allow for its commodification and exchange.

It has been well established that the coordinate data provided by services like *Twitter* and *Foursquare* are not nearly as accurate as their decimal points suggest (Xu et al. 2012). For example, a *Twitter* user may set their location to Boston, Massachusetts, while the user has specified their address to the city level, geocoding these results produces a location specified to within a single meter. While potentially misleading, this issue is well known and researchers have proposed solutions (Hecht *et al.* 2011). Check-in information, the basis of most LBSNs like *Foursquare*, avoids this potential distortion by having end-users check into geocoded venues. In order to encourage use, *Foursquare* and other services offer rewards for checking into certain venues. The first time a user checks into a restaurant they may be offered discount on their meal or a free drink from the bar. Similarly, a customer may receive a permanent discount after checking into a location a certain number of times. By rewarding user participation, *Foursquare* and other services incentive location fraud. Location fraud is to “falsely claim to be at a location, to receive undeserved rewards or social status” (Carbunar and Potharaju 2012, 1). Incentivizing end-users to contribute data, so that that data may itself be sold as a commodity, results in distorting end-user behavior to maximize the receipt of rewards.

For *Foursquare* and other LBSNs, this “fraud” doesn’t matter: the data provided is still valuable, they are capable of turning a profit, so whether an individual user is actually at a venue matters less than that the individual user *claims* to be at said venue. In fact, *Foursquare* allows and encourages applications like *Check In Take Out* which allow users to check into distant restaurants in order to place take-out orders. Figure 2 shows *Check In Take Out*’s use of *Foursquare*’s basic map presentation. So long as the data produced can be exchanged for a profit, the LBSN has no incentive to prevent this separation of physical location from check-in “location.” From the perspective of the end-user, whether they are actually at a check-in location matters less than the rewards they receive for setting their “location” to the venue. Where an end-user is physically located is less important than the quantified data representing “location.” While this has led some technology writers to repeatedly proclaim the death of the check-in (Mitchell 2012), it can more generally be seen as part of a shift in focus of mobile applications from simply recording location and providing destinations to shaping consumption patterns of users (Thatcher 2013; Wilson 2012).

This framework of data generation is driven by an innate profit motive – end-users receive discounts, *Foursquare* sells data to partners, partners use data to drive consumption – that is completely divorced from an accurate representation of physical location. Researchers who make use of this data are inherently accepting this framing for their research. Abductive exploration of Big Data may reveal patterns, but it reveals patterns of “location” as a commodity. Movement patterns found within *Foursquare* information reflect movement first encouraged and then shaped by motives for profit. A behavioral loop is created for both the end-user and those who study them: “A person feeds in data, which is collected by an algorithm that then presents the user with choices, thus steering behavior” (Lohr 2012).

Along with the opportunities offered by massive spatial data sets comes a set of restrictions both on how data can be accessed and what can be known through it. Spatial Big Data currently involves data created and shaped by a motive for profit. While this does not necessarily matter for a start-up company, it has profound repercussions for academic researchers. The next section offers a potential means of addressing the issues at stake in the research of spatial Big Data. It suggests looking to a series of parallel debates that occurred within the GIS research community throughout the 1990s as a means of deepening theoretical critiques and understanding of Big Data.

#### **4. Conclusion: Big Data and GIS**

With upwards of 80% of all data stored by businesses and governments having a spatial component, it has been suggested that GIS scholars hold a “home field advantage” when it comes to the study of Big Data (Farmer and Pozdnoukhov 2012). In addition to expertise in the handling and analysis of large, spatially-referenced data sets, the debates which have and continue to occur within the GIS research community speak directly to the concerns raised in the previous sections.

First, the data sets researchers rely upon are increasingly generated through and controlled by privately held corporations. This raises distinct concerns on how what can be known has become regulated by a small set of corporate entities. This parallels discussions of a rising technocracy controlling knowledge production in GIS. “This technocracy is hidden in the offices of the vendors that develop the hardware and software and make the technology more generally accessible” (Obermeyer 1995, 78). As GIS technology advanced, the mediation by technology receded from active consideration. Like the telephone, GIS became a “transparent” technology used without conscious consideration of the underlying technical processes (*ibid.*, 81). A similar situation has arisen for Big Data researchers: APIs standardize the



process of access, structuring the data, but they also set the limits of what the data contains. Burgess and Bruns (2012) have demonstrated the very heuristics of their analysis are shaped by the format of the data available through the API used. *Foursquare's* recent API change demonstrates that the very content of the data, and therefore what can be known through it, are subject to regulation and alteration outside of the control of researchers. The opaque process of corporate decision-making serves as the hidden technocracy of Big Data.

Finally, the reliance upon data generated with an explicit motive for profit – both for the end-user and the corporation – results in epistemological commitments not dissimilar to concerns raised with regards to the knowledges and approaches privileged by GIS use. For GIS, and now for Big Data, there is a need to distinguish between “empirical and technical claims about objects, practices, and institutions,” the discourses within which these claims, and claims to truth, are made (Pickles 1995, 23). Big Data, like GIS, accepts that a certain quantitative representation of life can stand in for its full meaning (Curry 1997). Further, in this inscription of meaning, Big Data must be seen as directly producing new knowledge rather than simply revealing it. The “hard work of theory” (Pickles 1997, 370) will tie Big Data directly to much longer traditions of social theory and social thought as they have engaged technology. Cartographers and GIS researchers have drawn productively from Benjamin (Kingsbury and Jones 2009), Heidegger (Pickles 1995), Foucault (Harley 1989) and other classical social theory perspectives in deepening an understanding of the relation between the specific technological form and the knowledges produced.

This paper has outlined some of the concerns with an increased reliance upon Data Fumes. Drawing from earlier ‘systems vs. science’ debates, and leveraging their ‘home field advantage,’ Cartographers and GIS researchers are in a unique position to contribute to the emerging field of Big Data Research. What can be put ‘on the map’ with spatial Big Data is now a product of, first, what is captured and, then, what data is made available by its owners – a group often distinct from who actually produced the data. Rather than letting these numbers ‘speak for themselves,’ Cartographers and GIS researchers should draw from their own long history to ground spatial Big Data and its practices in a more nuanced understanding of visualization, representation, and power.

## References

- Baker S (2013) Can Social Media Sell Soap? *The New York Times*. <http://www.nytimes.com/2013/01/06/opinion/sunday/can-social-media-sell-soap.html?hp&r=1&> Accessed 2 Feb. 2013
- Batty M (2012) Smart cities, Big Data. *Environment and Planning B* 39:191-193
- Benner J & Robles C (2012) Trending on *Foursquare*: Examining the Location and Categories of Venues that Trend in Three Cities. *Proceedings of the Workshop on GIScience in the Big Data Age 2012*: 27-35
- Berry DM (2011) *The Philosophy of Software: Code and Mediation in the Digital Age*. London: Palgrave Macmillan
- boyd D & Crawford K (2012) Critical Questions for Big Data. *Information, Communication & Society* 15(5): 662-679
- Brownlee J (2012) This Creepy App Isn't Just Stalking Women Without Their Knowledge, It's a Wake-Up Call About *Facebook* Privacy. *Cult of Mac*. <http://www.cultofmac.com/157641/this-creepy-app-isnt-just-stalking-women-without-their-knowledge-its-a-wake-up-call-about-Facebook-privacy/> Accessed 2 Feb. 2013
- Burgess J & Bruns A (2012) *Twitter* Archives and the Challenges of "Big Social Data" for Media and Communication Research. *M/C Journal* 15(5)
- Carbunar B & Potharaju R (2012) You Unlocked the Mr. Everest Badge on *Foursquare*! Countering Location Fraud in Geosocial Networks. *Proceedings of the 9th IEEE International Conference on MASS*
- Cerrato P (2012) Big Data Analytics: Where's the ROI? *InformationWeek: Healthcare*. <http://www.informationweek.com/healthcare/clinical-systems/big-data-analytics-wheres-the-roi/240012701> Accessed 2 Feb. 2013
- Curry M (1997) The digital individual and the private realm. *Annals of the AAG* 87: 681-699

Farmer C & Pozdnoukhov A (2012) Building streaming GIScience from context, theory, and intelligence. *Proceedings of the Workshop on GIScience in the Big Data Age 2012*: 5-10

Harley J (1989) Deconstructing the map. *Cartographical* 26: 1-20

Hecht B, Hong L, Suh B, & Chi E (2011) Tweets from Justin Bieber's heart. *Proceedings of the ACM CHI Conference 2011*

Hey T, Tansley S, & Toelle K (eds) (2009). *The fourth paradigm: Data-intensive scientific discovery*. Richmond, WA: Microsoft Research

Jacobs A (2009) The Pathologies of Big Data. *ACM Queue* 7(6): 10-22

Kitchin R, Dodge M (2007) Rethinking maps. *Progress in Human Geography* 31: 331-344

Laurila J, Gatica-Perez D, Aad I, Blom J, Bornet O, Do T, Dousse O, Eberle J, Miettinen M (2012) The Mobile Big Data Challenge. *Nokia Research*. [http://research.nokia.com/files/public/MDC2012\\_Overview\\_LaurilaGaticaPerezEtAl.pdf](http://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf) Accessed 2 Feb. 2013

Lohr S (2012) Sure, Big Data Is Great. But So Is Intuition. *The New York Times* [http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html?\\_r=3&adxnnl=1&partner=rss&emc=rss&adxnnlx=1357590814-L6vdMVi0JnNF0dB5hk1KLg&](http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html?_r=3&adxnnl=1&partner=rss&emc=rss&adxnnlx=1357590814-L6vdMVi0JnNF0dB5hk1KLg&) Accessed 2 Feb. 2013

Long X, Jin L, & Joshi J (2012) Exploring Trajectory-Driven Local Geographic Topics in *Foursquare*. *Proceedings of ACM Ubicomp '12*: 927-934

Mitchell J (2012) Life After Death of the Check-in. *Read-Write*. [http://readwrite.com/2012/04/10/pronouncing\\_the\\_death\\_of\\_the\\_check-in](http://readwrite.com/2012/04/10/pronouncing_the_death_of_the_check-in) Accessed 2 Feb. 2013

NSF Press Release (2012) NSF Announces Interagency Progress on Administrative Big Data. [http://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=125610](http://www.nsf.gov/news/news_summ.jsp?cntn_id=125610) Accessed 2 Feb. 2013

Obermeyer N (1995) The Hidden GIS Technocracy. *Cartography and Geographic Information Science* 22(1): 78-83

Pickles J (1995) *Ground Truth*. New York: Guilford Press

Pickles J (1997) Tool or science? Gis, technoscience and the theoretical turn. *Annals of the AAG* 87: 363-372

Rasmus D (2012) Why Big Data Won't Make You Smart, Rich, Or Pretty. *Fast Company*. <http://www.fastcompany.com/1811441/why-big-data-won%E2%80%99t-make-you-smart-rich-or-pretty> 2 Feb. 2013

Thatcher J (2013) Avoiding the Ghetto through hope and fear: an analysis of immanent technology using ideal types. *Forthcoming in GeoJournal*.

Thompson C (2012) *Foursquare* alters API to eliminate apps like Girls Around Me. *About Foursquare*. <http://aboutFoursquare.com/Foursquare-api-change-girls-around-me/> Accessed 2 Feb. 2013

Wilson M (2012) Location-based services, conspicuous mobility, and the location-aware future. *Geoforum* 43(6): 1266-1275